Semantic Web for Earth and Environmental Terminology (SWEET)

Rob Raskin
NASA/ Jet Propulsion Laboratory (JPL)
California Institute of Technology

http://sweet.jpl.nasa.gov

What is the Semantic Web?

- Internet where:
  - Automated tools understand content of web pages
  - Search tools use semantics to aid search
  - Underlying knowledge base can be dynamically updated

Semantic Web – A Vision for Earth Sciences

- Enable automated tools to discover information, data, and knowledge from Earth science Web pages and data providers
  - XML tags placed around terms in Web pages, referencing defining ontologies
  - Search engines use domain knowledge inherent in ontlogies to improve search performance

Objectives

- Prototype a semantic web for Earth science data resources
  - Develop a collection of ontologies
  - Use ontologies to improve discovery of Earth science data
- Partner with GCMD
- Raise TRL from 2 to 3-4

Outline

- 1. Ontology Development
- 2. Experiences With Ontology Languages/Tools
- 3. Ontology-Assisted Search Tool
- 4. Software Agents
- 5. Follow-up Work and Conclusions

1. Ontology Development

What is an ontology?

- Many answers:
  - Formal world-view
  - Collection of terms and their interrelations
  - Formal representation of knowledge
  - Knowledge base used by automated tools to understand web resources
  - …
- Formal ontology can be expressed using XML language

Task: Build an Ontology Based Upon GCMD Keywords

GCMD Controlled keywords

- Earth science (~1,000 terms)
  - Example: EarthScience>Oceanography>SeaSurface>SeaSurfaceTemperature
- Instruments
- Missions
- Data Services
- Data Centers
- …

GCMD Uncontrolled keywords

- ~20,000 terms submitted by data providers
  - Many are abstract (climatology, surface, El Nino, EOSDIS)

Ontology Development Strategy

- We focused on developing an ontology structure for Earth science concepts
  - Based on (but not limited by) GCMD science keywords
  - "Faceted" approach produced orthogonal keyword structure (somewhat different from GCMD)
- Assumed that higher level concepts (e.g. PhysicalQuantity, AbstractThing) were defined in other ontologies and could be imported

SWEET Science Ontologies

- EarthRealms
  - Atmosphere, SolidEarth, Ocean, LandSurface, …
- PhysicalProperties
  - temperature, composition, area, albedo, …
- NonLivingSubstances
  - CO2, water, lava, salt, hydrogen, pollutants, …
- LivingSubstances
  - Humans, fish, …

SWEET Conceptual Ontologies

- Phenomena
  - ElNino, Volcano, Thunderstorm, Deforestation, Terrorism…
  - Each has associated EarthRealms, PhysicalProperties, spatial/temporal extent, etc.
  - Specific instances of phenomena can be defined: e.g., 1997-98 ElNino
- HumanActivities
  - Fisheries, IndustrialProcessing, Economics

SWEET Numerical Ontologies

- SpatialEntities
  - Extents: country, Antarctica, equator, inlet, …
  - Relations: above, northOf, …
- TemporalEntities
  - Extents: duration, century, season, …
  - Relations: after, before, …
- NumericalEntities
  - Extents: interval, point, 0, positiveIntegers, …
  - Relations: lessThan, greaterThan, …

SWEET Other Ontologies

- Units
  - Extracted from Unidata's UDUnits
  - Added SI prefixes (km is a type of m with an associated conversion factor of 1000)
- DatasetProperties
  - Extracted from GCMD, XDF, ESML
- WebServices
  - Extracted from GCMD; other services added

Desirable Ontology Properties

- Scalability
  - Easily extendable to enable specialized domains to build upon more general ontologies
- Orthogonality
  - Compound concepts decomposed into their component parts, to make it easy to recombine concepts in new ways
- Community developed
  - Community input should guide development

Desirable Ontology Properties *(cont.)*

- Language independence
  - Representation of *concepts*, rather than terms. Concepts independent of slang, technical jargon, foreign languages
  - Synonymous terms (e.g., marine, ocean, sea, oceanography, ocean science) can be mapped separately to an ontology element
- Application independence
  - Ontology structure and contents based upon inherent knowledge of discipline, rather than on how knowledge is used

2. Experiences With Ontology Languages and Tools

Ontology Languages

- RDF
  - Specialization of XML
  - Standardizes basic concepts:
    - Class, subclass, property, subproperty, domain, range, imports, …
    - Simliar to how <b> and <p> are standardized in HTML
  - Parsing tools widely available

Ontology Languages (cont.)

- DARPA Markup Language + Ontology Inference Language (DAML+OIL)
  - Specialization of RDF
    - Adds: cardinality, transitive & inverse properties, …
  - Enables ontology interoperability, extendibility, reusability
  - Cyc and open source subset "OpenCyc" (largest existing ontologies) have been translated into DAML
  - Adopted for this project
    - Enables use of higher-level concepts defined elsewhere
- Ontology for the Web Language (OWL)
  - Version of DAML+OIL being adopted by W3C as official standard

DAML and Numbers

- DAML has minimal support for numbers
- Numeric objects defined only through an XSD spec
  - Real interval and sequence of integers can be defined and extended (although awkwardly)
- Numeric operators not defined at all
  - No operators for: max, greaterThan, overlap, …
  - Major deficiency, as many science concepts are defined numerically

DAML and Numbers
(Example: Definition of visible light)

```
<daml:class rdf:ID="VisibleLight">
    <rdfs:subclassOf> ElectromagneticRadiation </rdfs:subclassOf>
    <rdfs:sameClassAs>
        <daml:restriction>
            <daml:onProperty rdf:resource="#Wavelength" />
            <daml:toClass daml:class="Interval300to800" />
        </daml:restriction>
    </rdfs:sameClassAs>
</daml:class>
```

Difficulties:

- Class "Interval300to800" must be separately defined!
- A property "moreEnergetic" is desirable. It is isomorphic to the "lessThan" relation on the real numbers (but "lessThan" is not defined in DAML)

DAML and Numbers (Example: Space and Time)

- Most spatial and time concepts are easily mapped to numeric relations.
- Spatial concepts require definition of 2-D or 3-D numerical system
  - Cartesian product of the real line had to be defined
- No relevant space/time ontologies were available a priori
  - Gazetteers limited to bounding box of region
  - Temporal concepts not specialization of numeric concepts

Dimensions
(lat, lon, vertical, and time are orthogonal)

	Example objects	**Example relations**
Space (3-D)	Africa, Pacific Ocean	1-D boolean valued: west, south, above 2-D space valued: surface, floor
Time	Day, Season, Moment	before, after, during

Database Storage

- XML-based languages (e.g. DAML) useful for data/model exchange; not very practical for *storage* and *query* of large ontologies
  - DBMS is highly desirable
- Postgres object-oriented DBMS
  - Stores class names and parent relations
  - 2-way translation tools developed between XML and database representations

DAML/OWL Tools

- Editing/visualization tools very limited
- OilEd ontology editor does not support all features of DAML (e.g. derived number types)
- Database API would be helpful

3. Ontology-Assisted Search Tool

Prototype Search Tool

- Search tool finds additional terms that are likely to match search
  - Synonyms
  - Parent concepts
- Submits union of these terms to another search engine (GCMD Search tool)

Example: Phenomena

- El Nino defined in terms of its facets
  - Earth Realm (oceanography, atmospheric science)
  - Physical Property (wind, temperature, pressure, precipitation, …)
  - Spatial Extent (tropical Pacific)
- Specfic El Nino events defined as instances of this phenomenon
  - E.g. 1997-98 El Nino has associated spatial and temporal extents

How to get OWL tags onto web pages?

- Will Web page creators voluntarily place ontology tags on their Web pages?
- Tags can be virtually inserted during the indexing process
  - Requires tools from natural language processing to interpret text and classify (cluster) alternate meanings of words

4. Software Agents

Range of Query Types

1. Data Specific: Get ASTER data for Ecuador from Dec 7, 2000 to Dec 31, 2000
   - ASTER (Instrument)
   - Ecuador (SpatialEntity)
   - Dates (Time)
2. Researcher oriented, non-instrument specific: Visible and near infrared data, high resolution, of Duke University forest (lat, long provided as bounding box), for Jan 1971-present
   - Visible, near-infrared (ElectromagneticRadiation)
   - Latitude/longitude (SpatialEntity)
   - Resolution (inferred to be spatial)
   - High (Quantities)
   - Dates (Time)

Range of Query Types (cont.)

3. Educated public request: show me data from 1993 ENSO
   - ENSO (Phenomena lists SameAs ElNino and gives associated quantities:)
     - EarthRealms (Oceanography, Weather)
     - PhysicalProperties (Temperature, Moisture)
     - SpatialRegions (Tropics)
   - 1993 (Timeline gives specific ElNino time extent)
4. Public request: show me my house
   - My (Agent interprets word as specific to user, prompting more information: City, State, Country)
   - House (HumanActivities)
   - Show (WebServices – request for image)

Agent Network

- "Request Agent" is defined for each of these four request classes
  - Request agents complete subtask then hand problem to next request agent
  - Elicits more information when needed
- Ontology agent is defined for each ontology
  - Agents lookup requested terms in the ontology and infer nature of request

5. Follow-up Work and Conclusions

Federation SEEDS Prototype *(funded)*

- Incorporate Search agent in ESIP Federation Search Page
- Include indexing of all Federation pages
- Enable searchers to access data products (in addition to web pages)

SEEDS Phase 2 *(funded)*

- Work closely with other infrastructure initiatives to create a common semantic framework
  - ESML, GCMD, ESMF, OGC, IPG, Geoinformatics, etc.
- Improve spatial/gazetteer support
  - Represent countries and features as polygons
- Expand work on search tool agents

Contributions of SWEET

- Improved data discovery without exact keyword matches
- TRL advanced from 2 to 3
- SWEET Earth Science, spatial, temporal, and numeric ontologies will be submitted as contributed DAML libraries
  - Domain specialists can specialize our work
  - Space, numerics, and event ontologies will have a general appeal

Conclusions

- We are in the early stages of knowledge representation
- Languages and tools could be more robust
- There is wide range of potential applications that could take advantage of ontologies
- SWEET is a starting point for representation of knowledge in Earth sciences
- Agent tools benefit from segmenting users into "user types"
- Others can extend what we have developed

Contacts

- SWEET http://sweet.jpl.nasa.gov
- Rob Raskin raskin@jpl.nasa.gov